# Statistical Analysis on the Recklessness of Young Drivers in Hong Kong

Chun Ki Yip- 1004532589

April 16, 2021

## Abstract

<Here you should put a summary of the entire report. You should write the section at the end (i.e., after finishing the other 5 main sections).>

## Introduction

Young drivers are commonly stereotyped to be reckless and causes more car accidents than any age group. Young drivers recieve discrimination from insurance companies, car rental companies and from examiners of the driving tests. Personally, there are countless of times when I rent a car and the rental fees go from 30 dollars to 200 dollars because of young fees and extra insurance fees, it frustrates that someone with a perfect driving record is punished because of age. This report will examine all the private car accidents in the wonderful city of Hong Kong from 2014 to 2019 and analyze using various statistical methods such as; bootstrapping confidence interval, Bayesian credible interval, maximum likelihood estimation, liner regression, hypothesis test and goodness of fit test to see whether or not younger drivers create more car accidents or have a higher tendancy to be involved in a serious or fatal accident.

The data we are looking at are from the package "Datasets Related to Hong Kong" [], which contains election data and data related to car accidents. The data we will be focusing on today are the car accident data set which the package pulls directly from the official website of the Transport Department of Hong Kong. This data provides a complete record of every traffic accident for the past 5 years which is very valuable for analysis, using this data we could extend this anlaysis to other big metropolis globally.

Some terminologies that would be used in this report are Beyesian inferenece, which is a statistical methodology that uses data and families of distributions to estimate the parameters of a distribution which best fits the data with the Bayes theorem []. Maximum likelihood estimatior is a methodology that gives a way to pick the highes liklihood distribution function that could describe the data distribution []. Bootstrapping confidence interval is a methodology that requires constant sampling and replacement of a sample population to create more data points to give a more accurate range of where the sample statistics truely lies in []. Finally, Bayes's theorem is a mathematical statement that describes the relationship between prior conditions and the probability of the event as the following: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Hypothesis 1: Drivers of all ages that are involved in a car accident have the same tendency of having serious/fatal car accident.

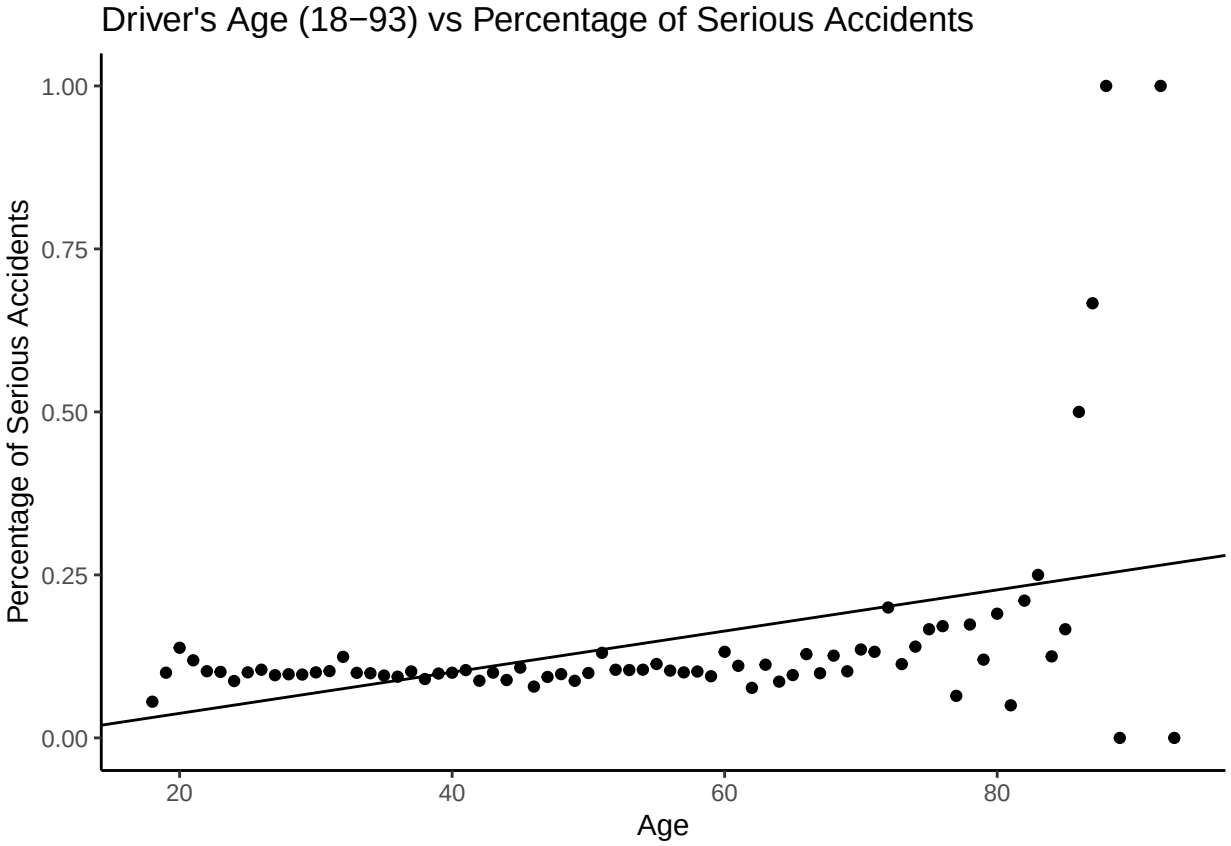
Hypothesis 2: Age and frequency of car accidents of the age has no relationship

## Data

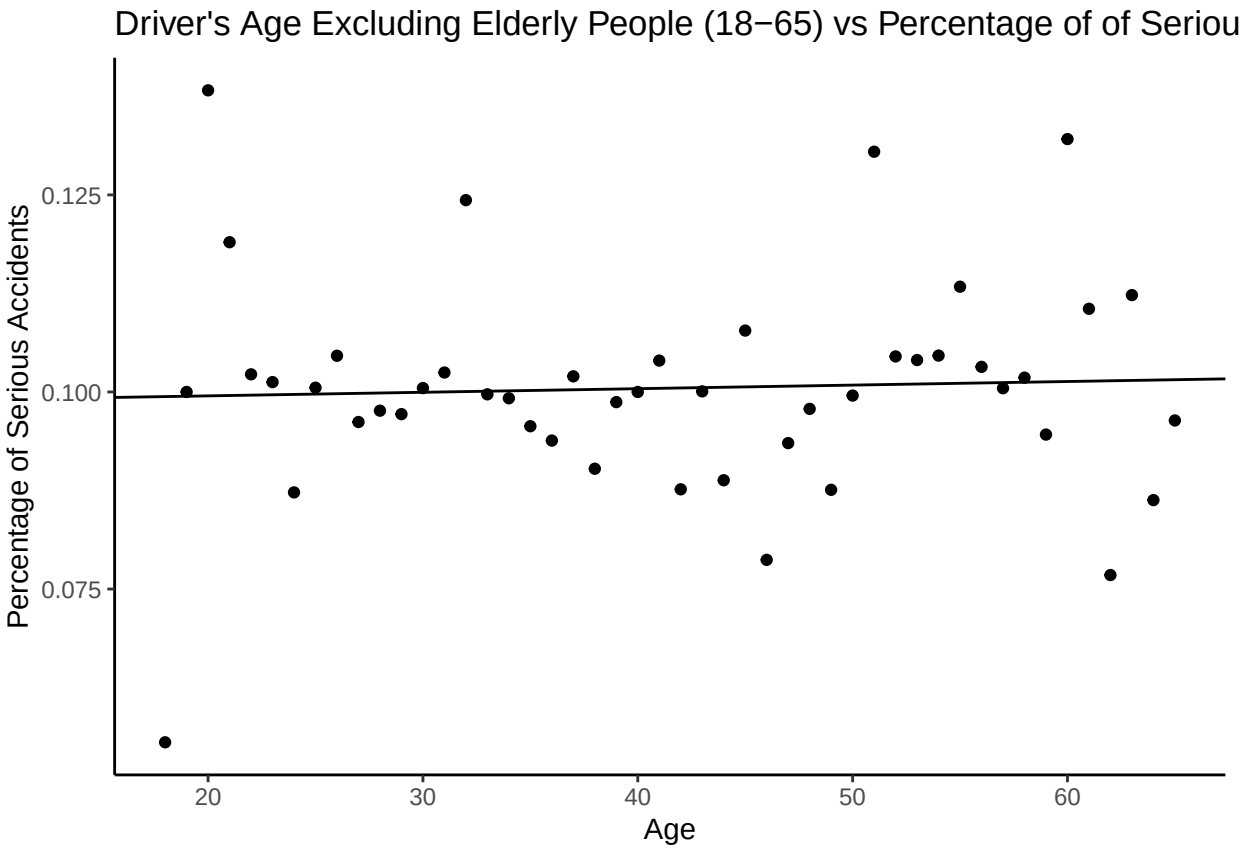
The vehicle data of the hkdataset includes a complete record of every car accident that occurred during 2014 to 2019 in Hong Kong. Some variables in the data set are the year of the accident, gender, manufacture year

of the car, driver's age, severity of the accident in three levels; slight, serious, and fatal, and the type of vehicle. The important variables we will be looking at are the driver's age and severity of the accident. The report will also focus on only the private car accidents because other vehicle types such as trams, taxi, and trucks lack the representation of young drivers which we are most interested in. Each data point is collected individually by Hong Kong traffic officiers when the accident occur and then transferred to the Transport Department website [].

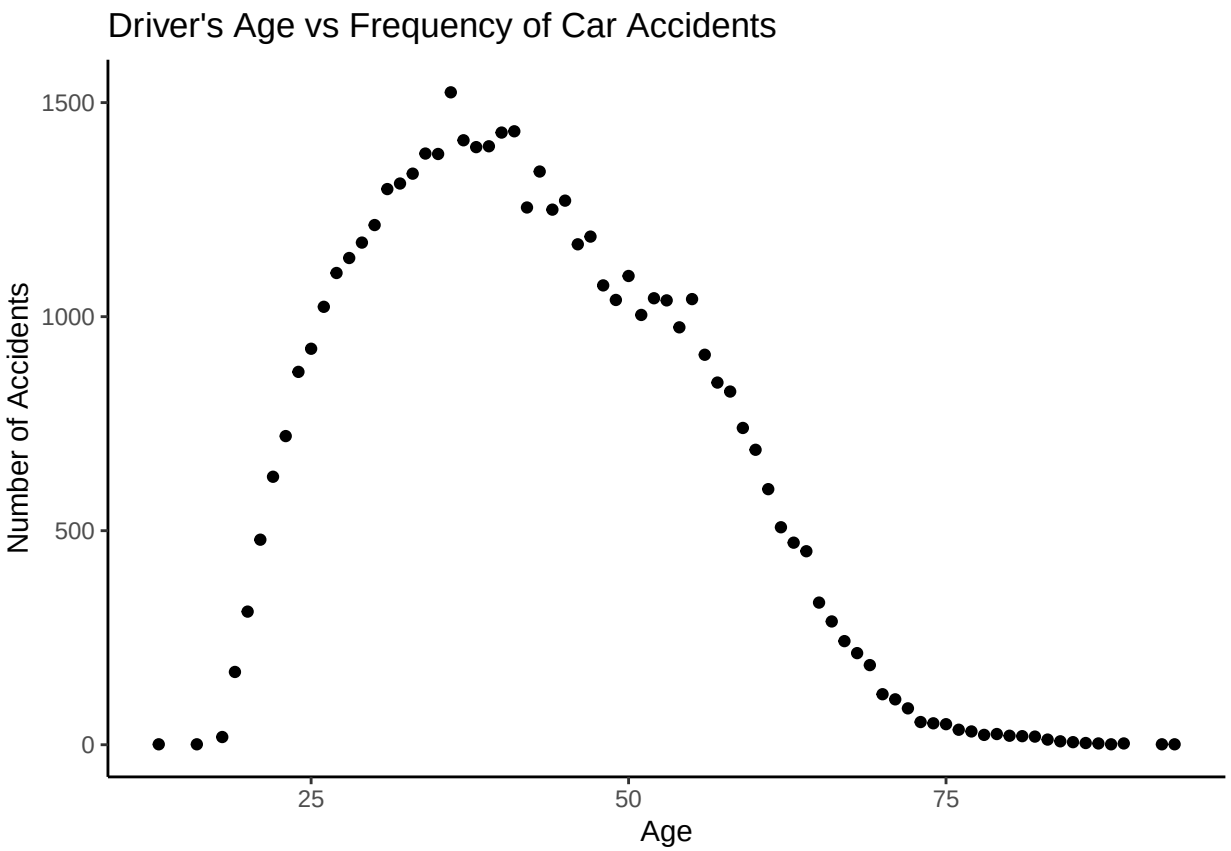
The cleaning process of the data started with selecting and filtering data points that are needed for the analysis using functions from the tidyverse package []. The group_by function and pivot_wider function in the tidyverse package was used to group the data by age and severity level, and then the percentage of serious and fatal injuries by age was calculated after tallying the frequencies of accidents of each severity level in every age []. I chose to do a lot of my analysis on the percentage of serious and fatal accidents rather than just the total accidents because the total number accidents reflect more of the population of drivers of each age instead of the recklessness of the driver which we are interested in, using the ratio of serious and fatal accidents over all car accidents gives every age group a fairer outlook as not every age group is proportioned equally.




This is a scatter plot created using the tidyverse package [] shows the driver's age against the percentage of serious and fatal accidents, this graph includes ages of 18-93 which already excludes some data points with less than 10 accidents. T simple linear regression line has a slope of 0.0031564 which shows that there is a positive relationship between age and percentage of accidents which doesn't support our hypothesis.




This is a scatter plot created using the tidyverse package [] graphs the age against the percentage of serious and fatal accidents of ages 18 to 65. This grpah excludes the elderly people above age of 65 because from the previous graph we can observe that a couple of data points from the elderly people are significantly higher than most of the other which made the linear regression line a lot steeper, the high fatality rate of elderly people in car accidents could have factors like other health conditions which skews the statistics After removing those points we can see that the slope is rather flat, although still positive with a slope of 4.559e-05, this is interesting because the linear regression supports that there are no difference in the percentage of serious and fatal accidents from younger drivers to older drivers which stereotypes suggests. Although this observation is interesting, the rest of the report would continue to use the data which includes the elderly people for completeness.




This is a scatter plot created using the tidyverse package [] graphs the age of driver against the number of accidents caused by them. This graph shows that distribution of data like the normal distribution which suggests that instead of a relationship between age and frequency of accidents this may show snapshot of the overall age distribution of drivers in Hong Kong. The data is skewed towards the left side meaning there are a denser population of younger drivers compared to old drivers.

Table 1: Data Summary Table

| Type of Data | Mean | Median | Max | Min | Standard Deviation | Variance |
|---|---|---|---|---|---|---|
| Percentages of Serious Accidents | 0.146701888252691 | 0.102193479679824 | 1 | 0 | 0.167636588352746 | 0.028102025754548 |
| Age of Driver involed in Accident | 42.1423772556894 | 41 | 93 | 18 | 12.4109951580485 | 154.032800813103 |

This data summary table shows the critical data summary of two variables of the data; the percentage of serious accidents, and the age of the driver involved in the accident. This data table is displayed using the knitr package.

All analysis for this report was programmed using `R version 4.0.4`.

# Methods

This section would introduce the methodology of the statistical tests that was conducted for this analysis including the linear regression, confidence interval, maximum likelihood estimator, hypothesis test, goodness of fit test, and Bayesian credible interval.

### Linear Regression

The two linear regression are used to show the relationship/lack of relationship in the Age vs Percentage of Serious/Fatal Accidents graphs. Both of them shows a very small positive slope meaning the two is likely to have no relationship. The linear regression model made using the lm function in R.

### Confidence Interval

The CI we're interested in is the 95% CI of the percentage of serious/fatal accidents divided by age groups of below 24 which represents the young drivers, then going up by intervals of ages of 9 from 25-34, 35-44, 45-54, 55-64 and lastly 65+. These age group was derived from first dividing the young drivers which I put the cutoff at 18-24 we are interested in and then dividing the rest of the ages equally without having too many categories. Having these interval of ages and getting the confidence interval of the percentage of serious/fatal accidents is useful because we can compare directly the tendencies serious accidents of each age group which relates to one of our hypothesis. To get our CI, I'm conducting a bootstrap of 10,000 repetitions each with 10,000 samples with replacement.

### Maximum Likelihood Estimator

In this MLE we assumed that the data follows a uniform distribution, which is a reasonable choice looking at the flat linear regression from the graphs earlier. I have used the MLE approach to estimate that the intervals of the data $\theta_1$ and $\theta_2$ are $min(x_1, x_2, ..., x_n)$ and $ max(x_1, x_2,...,x_n)$ respectively. All derivations regarding the MLE can be found in Section 1 of the Appendix.

### Hypothesis Test

A hypothesis test on whether the average age of a driver involved in a car accident was 49 years old, which was an assumpmtion and generous estimate of the age of drivers by averaging the ages of drivers without the outliers. Another assumption was made on the data being a normal distribution since we have a very large sample size of 48821 data points, and the p-value was found using the Z method and compared to $\alpha = 0.05$ to reject or accept the null hypothesis.

### Goodness of Fit Test

I am using the statistical test $-2log\frac{L(p_o)}{L(\hat{p})}$ to find if the frequency of accidents of ages from 18-25, 26-65, and 65-93 are distributed equally after being weighted by the range of the age group. Here I am assuming that each age contributes to same amount car accidents.

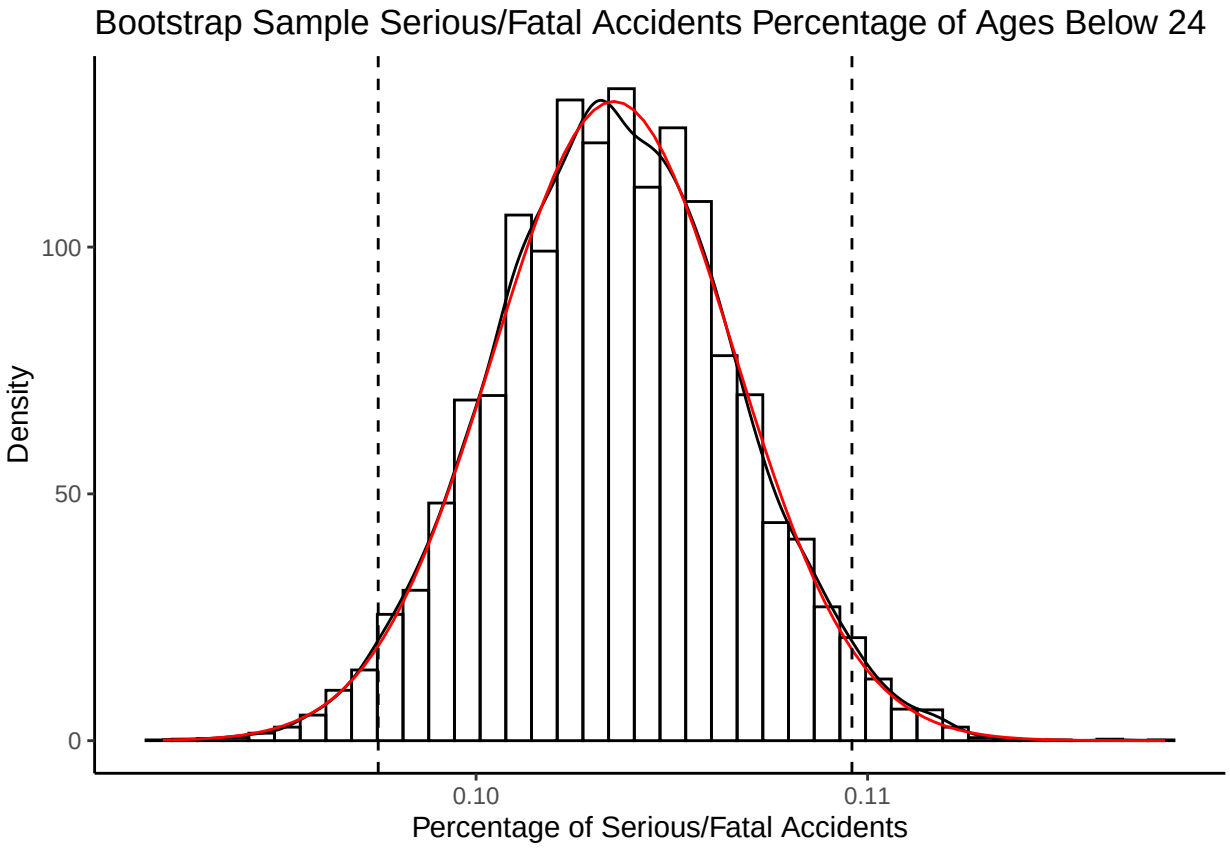
### Bayesian Credible Interval

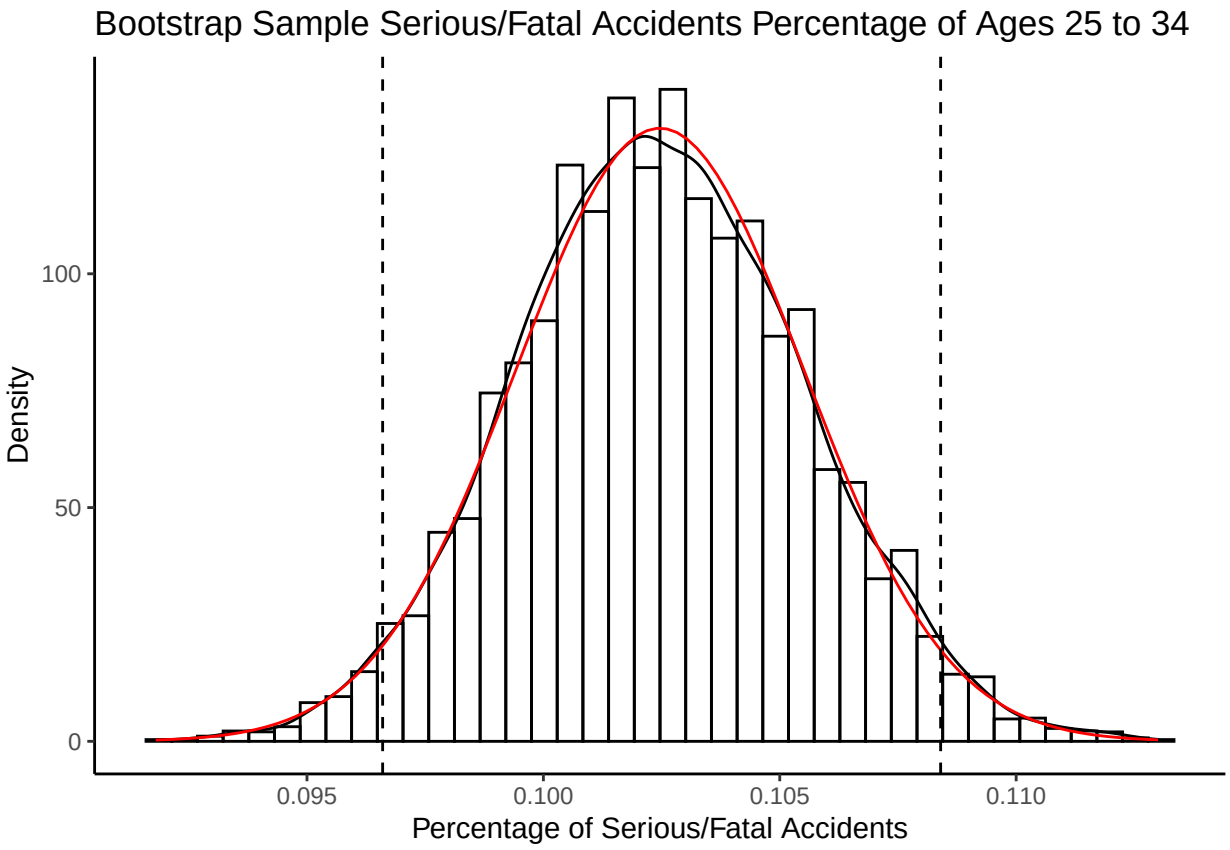
The posterior distribution derivation is where we derive a fitting distribution for the data set by using a prior distribution and the Bayes' Rule. The Bayesian Credible Interval methodology then uses the posterior distribution which we derived to get a confidence interval of a sample statistics. In this methodology I am assuming that the prior distriibution is a beta distribution with parameters beta(1,1) which resembles the flat linear regression of the data, and that the data is a random sample. I using the Bayesian Credible Interval to esitmate the CI of which 95% of the percentage of serious/fatal accidents lies in. All derivations regarding the posterior distribution can be found in Section 2 of the Appendix.

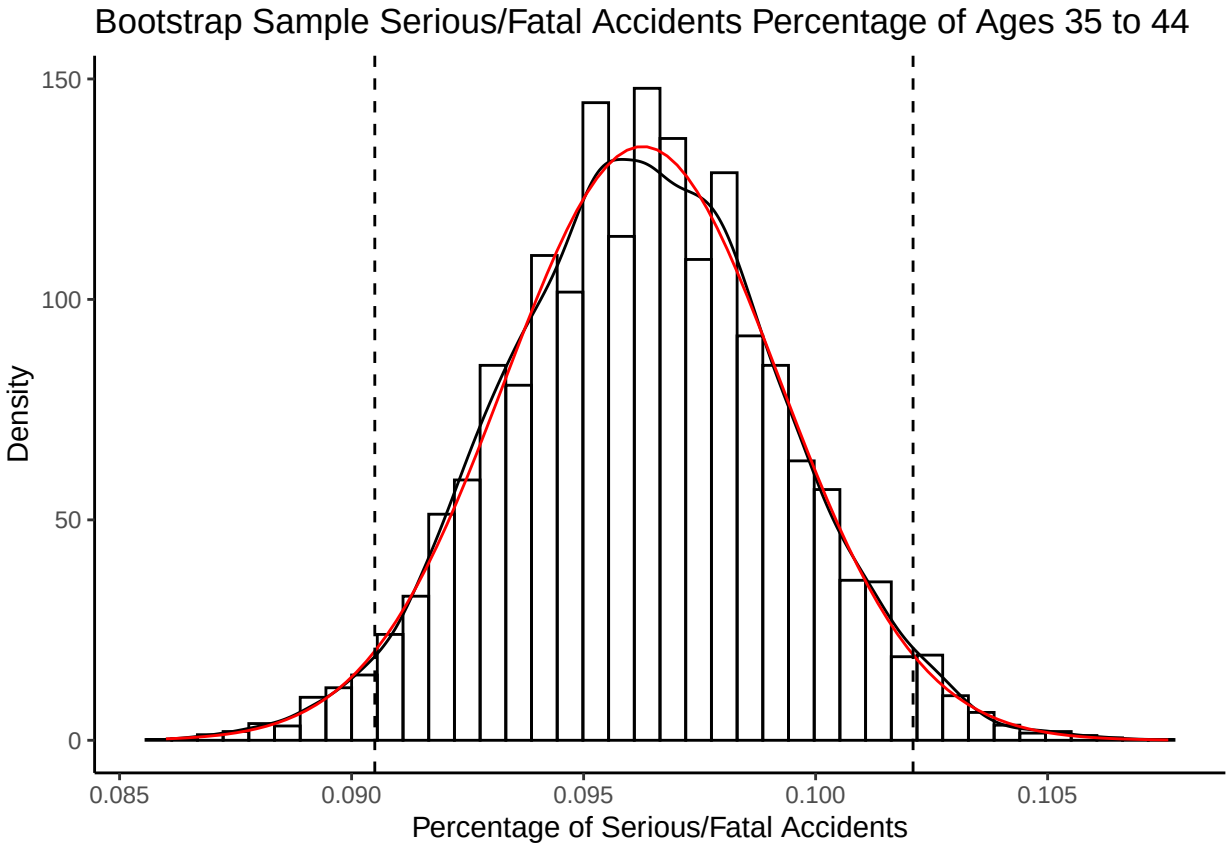
## Results

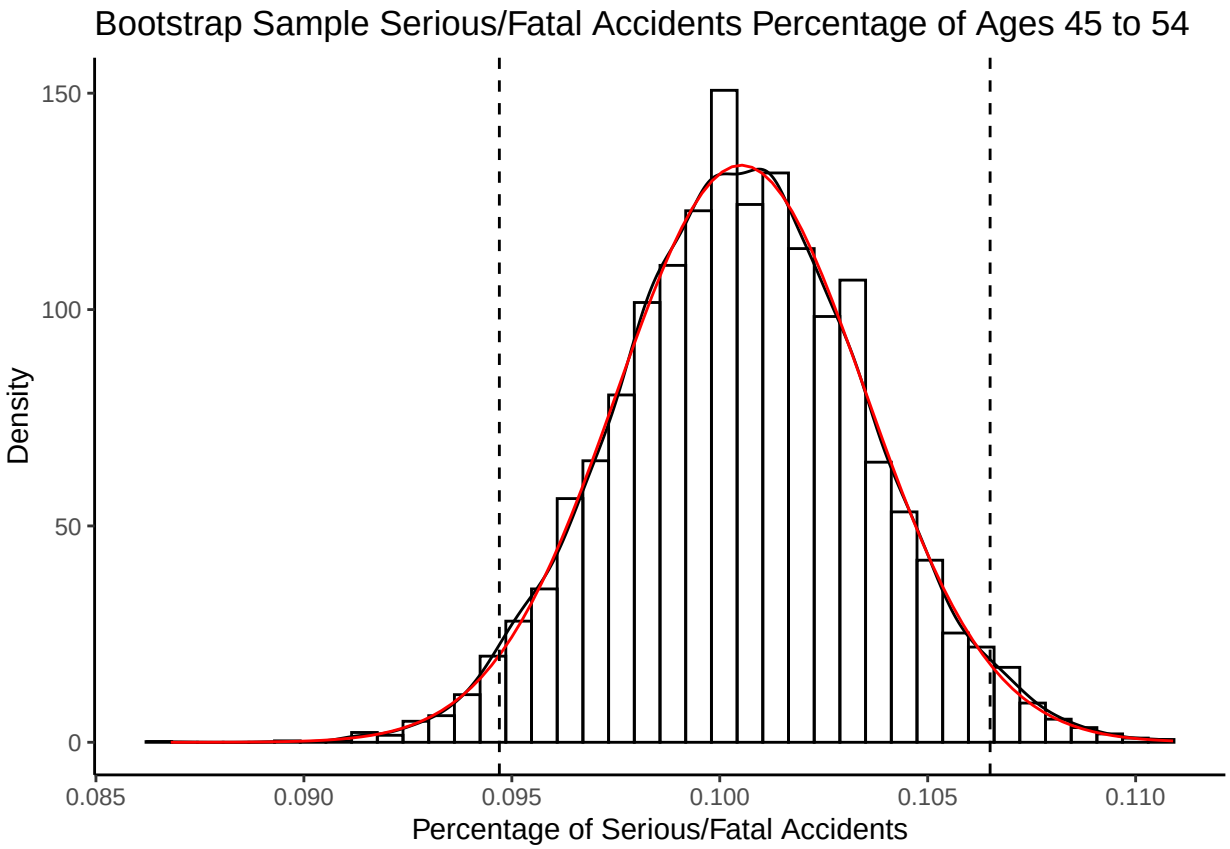
The statistical tests show disagreeing results, some suggesting that there are some slight negative or positive relationship between age and frequency of car accidents or age and the percentage of serious and fatal accidents while others suggests there are no relationship at all, overall, the relationship could be concluded as weak to none.

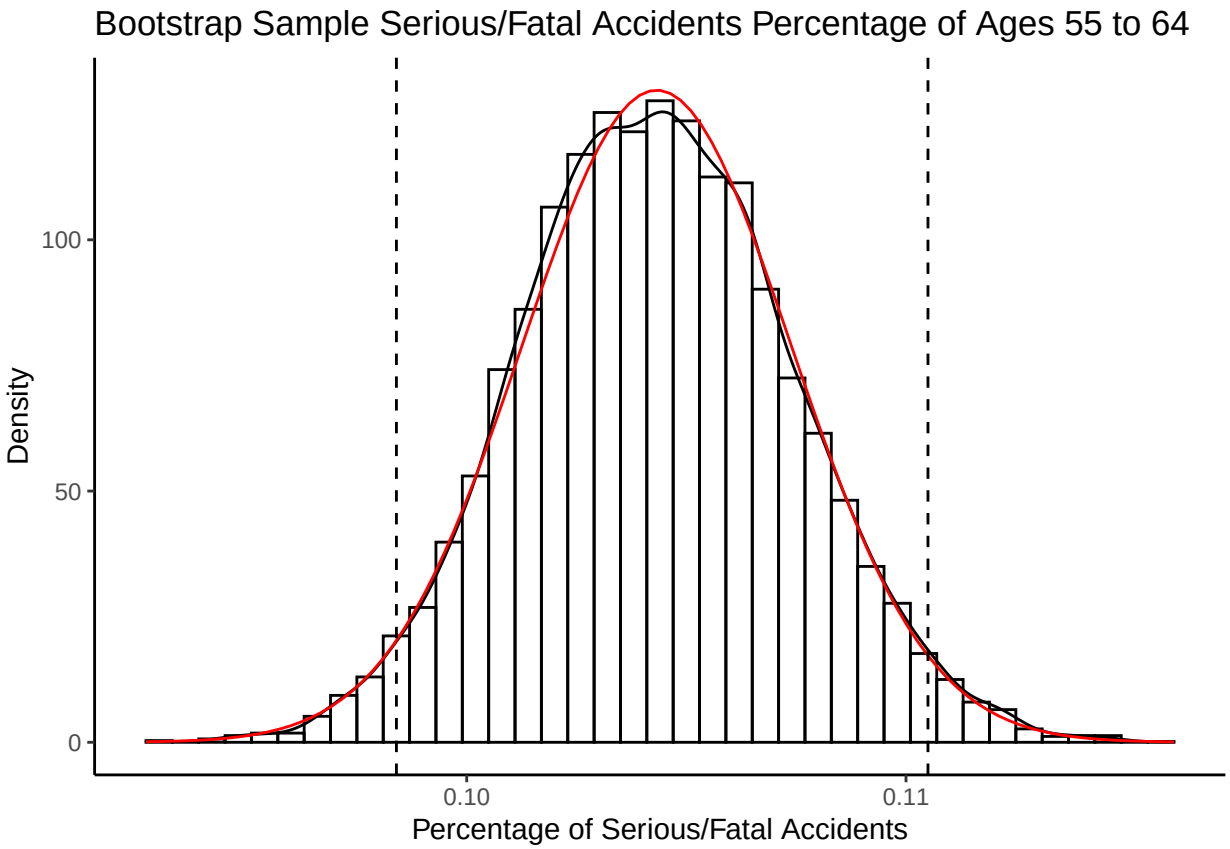
**Confidence Interval**

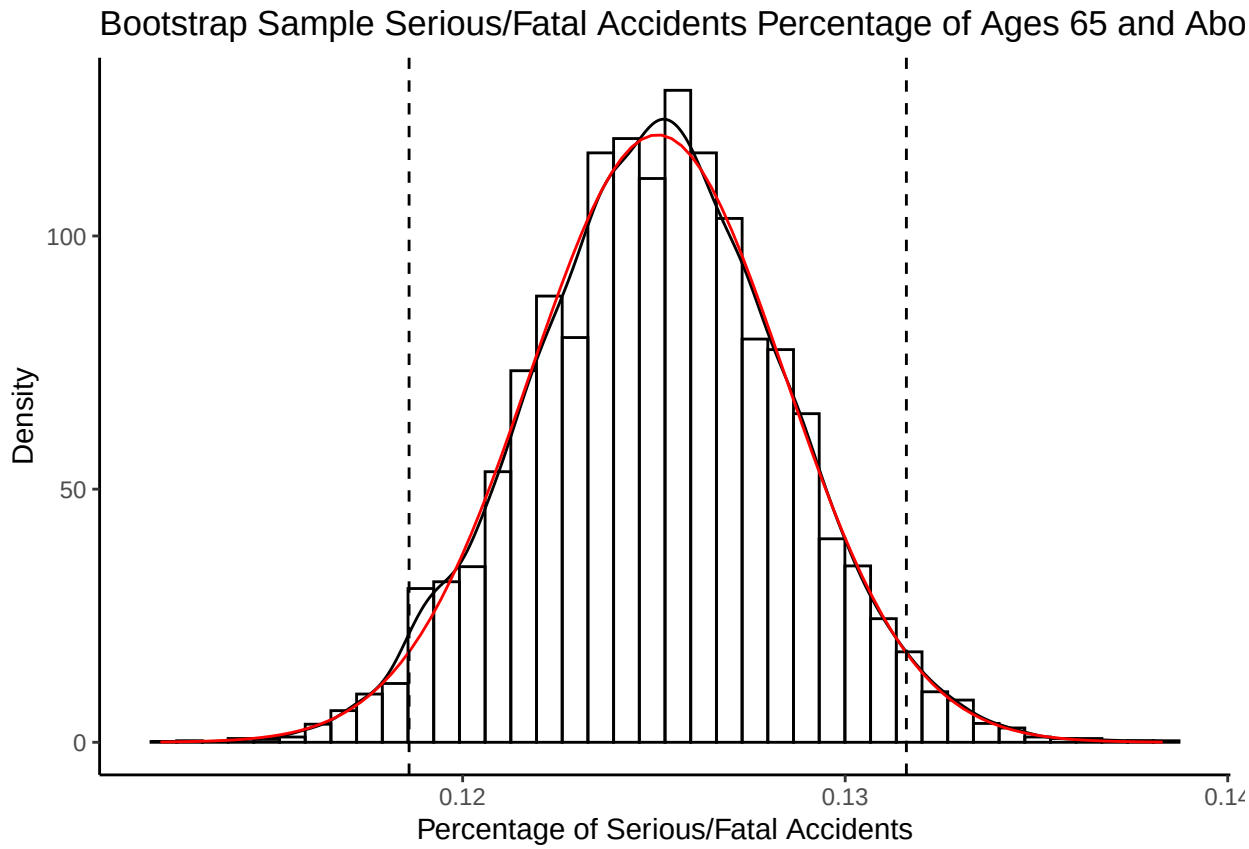



Bootstrap Sample Serious/Fatal Accidents Percentage of Ages 35 to 44
Density
150
100
50
0
0.085
0.090
0.095
0.100
0.105
Percentage of Serious/Fatal Accidents


Bootstrap Sample Serious/Fatal Accidents Percentage of Ages 45 to 54
Density
150
100
50
0
0.085
0.090
0.095
0.100
0.105
0.110
Percentage of Serious/Fatal Accidents


Bootstrap Sample Serious/Fatal Accidents Percentage of Ages 55 to 64
Density
100
50
0
0.10
0.11
Percentage of Serious/Fatal Accidents




These graphs show the boostrap distribution of different age groups, although they all look similar in shape, there are slight difference in the skew and range which is covers, this would be explained in detailed follwing the table below.

Table 2: Confidence Interval Table

| Age Group | Lower Bound | Upper Bound |
|---|---|---|
| Below 24 | 0.0975 | 0.1096000 |
| 25 to 34 | 0.0966 | 0.1084025 |
| 35 to 44 | 0.0905 | 0.1021000 |
| 45 to 54 | 0.0947 | 0.1065000 |
| 55 to 64 | 0.0984 | 0.1105000 |
| 65 and Above | 0.1186 | 0.1316000 |

This table compares the 95% CI of the percentage of serious/fatal accidents of each age group. Surprisingly, there was a decreasing trend from the ages 18 to 44, which is different from what the linear regression suggests. This supports that younger drivers have a higher tendency to be involved in a serious/fatal accident when they are involved in a car accident than older adults. Then as age goes up from 44, it shows a positive trend again, peaking the percentages at the group of 65+ which matches with the linear regression before. The fluctuation of percentages of serious and fatal accidents may suggests that there is no relationship at all and there are many more things that factors into a car accident.

### Maximum Likelihood Estimator

To get the maximum value of $\frac{1}{(\theta_2-\theta_1)^n}$ the value of $(\theta_2 - \theta_1)^n$ has to be minimized, meaning we should take $\hat{\theta}_1$ and $\hat{\theta}_2$ such that they are as close to each other as possible, i.e. take the largest possible $\hat{\theta}_1$ and smallest possible $\hat{\theta}_2$ keeping in mind that $\theta_1 \leq min(x_1, x_2, ..., x_n)$ and $\theta_2 \geq max(x_1, x_2, ..., x_n)$, therefore the largest possible $\hat{\theta}_{1MLE} = min(x_1, x_2, ..., x_n)$ and smallest possibles $\hat{\theta}_{2MLE} = max(x_1, x_2, ..., x_n)$.

### Hypothesis Test

Let's assume that the average age of the driver that was involved in an accident is 49 years old which is the average of 18 and 80, which is a generous estimate of the age range of most drivers active in Hong Kong.

Null hypothesis is $H_0 : \mu = 49$

Alternative hypothesis is $H_1 : \mu \neq 49$

Sample Statistics Mean = `mean(car_accidents$Driver_Age)`

Sample Statistics Standard Deviation = `sd(car_accidents$Driver_Age)`

Sample Size = `tally(car_accidents)`

Assume that the age of the drivers are normally distributed we get the following equation

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{42.1423772556891 - 49}{12.4109951580485/\sqrt{48821}}$$

The p-value= `pvalue` which is less than $\alpha = 0.05$ therefore we reject the null hypothesis, meaning our hypothesis of average age of 49 doesn't capture the true mean.

### Goodness of Fit Test

Goodness of Fit Test on Car Accidents Involving Young Adults (18-25), Adults (26-65) and Elderly People (66-93)

$H_0$ : Every category contribute to the equal amount of car accidents after being weighted or each age contributed to $\frac{rangeofagegroup}{rangeoftotalage}$ of the total car accidents.

$H_A$ : All categories do not have an equal amount of car accidents after being weighted or not all categories contributed to $\frac{rangeofagegroup}{rangeoftotalage}$ of the total car accidents.

Using the test statistics $-2log\frac{L(p_o)}{L(\hat{p})}$ we get that

We result in $L(p_0) =$ `p_0` and $L(\hat{p}) =$ `p_hat` which unfortunately could not be plugged into the test statistics because you can't take the log of 0, but we can make an inference that because the sample size of each category was too large, having a fraction take the power of this large number would result in an extremely small number which was rounded to 0 by R. Even if R could calculate this extremely small number, it would not make a good candidate for the goodness of fit test because the log of the small number would extrageate the result of the test.

### Bayesian Credible Interval

The 95% CI of percentage of serious/fatal accidents lie in the range of `BCI_lower_bound` and `BCI_upper_bound`

## Conclusions

The linear regressions suggests that there is no relationship between age and the tendency of serious or fatal accident especially when we compoare between young drivers of ages below 24 and adult drivers from ages 25

to 44, although elderly people show a increase in percentage of fatal and serious accidents, but it is hard to tell if it is due to other factors like health issues. The hypothesis test show that the average age of someone involving a an accident is not the same as the average age of drivers meaning that the disitrbution of total drivers of each age is either skewed left, or that younger drivers do indeed drive more recklessly, or it could be that drivers of age 44 drives recklessly. ## Weaknesses

A weakness in this report is there could've been more accurate assumptions of the prior distributions and distribution of the MLE.

### Next Steps

Look at different variables and factors that affect the percentage of serious and fatal accidents, and find more data that may sugggest or disagree with the current stereotype agianst younger drivers.

# Bibliography


1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

# Appendix

## Section 1

MLE for Driver's Age vs Percentage of Serious/Fatal Accidents Assume that $x_1, x_2, ..., x_n$ are $iid\ unif(\theta_1, \theta_2)$ $f(\theta) = \frac{1}{\theta_2 - \theta_1}$ $\theta_1 \leq x_i \leq \theta_2$ $L(\theta) = f_\theta(x_1, x_2, ..., x_n) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \cdots f_\theta(x_n) = (\frac{1}{\theta_2 - \theta_1})(\frac{1}{\theta_2 - \theta_1}) \cdots (\frac{1}{\theta_2 - \theta_1})$ s.t. $\forall x_i, \theta_1 \leq x_i \leq \theta_2$

$$\Rightarrow L(\theta) = \begin{cases} \frac{1}{(\theta_2 - \theta_1)^n} & \theta_1 \leq x_1 \leq \theta_2, \theta_1 \leq x_2 \leq \theta_2, ...\theta_1 \leq x_n \leq \theta_2 \\ 0 & otherwise \end{cases}$$

$$\Rightarrow L(\theta) = \begin{cases} \frac{1}{(\theta_2 - \theta_1)^n} & \theta_1 \leq min(x_1, x_2, ..., x_n), \theta_2 \geq max(x_1, x_2, ..., x_n) \\ 0 & \theta_1 > min(x_1, x_2, ..., x_n), \theta_2 < max(x_1, x_2, ..., x_n) \end{cases}$$

To get the maximum value of $\frac{1}{(\theta_2 - \theta_1)^n}$ the value of $(\theta_2 - \theta_1)^n$ has to be minimized, meaning we should take $\hat{\theta}_1$ and $\hat{\theta}_2$ such that they are as close to each other as possible, i.e. take the largest possible $\hat{\theta}_1$ and smallest possible $\hat{\theta}_2$ keeping in mind that $\theta_1 \leq min(x_1, x_2, ..., x_n)$ and $\theta_2 \geq max(x_1, x_2, ..., x_n)$, therefore the largest possible $\hat{\theta}_{1MLE} = min(x_1, x_2, ..., x_n)$ and smallest possibles $\hat{\theta}_{2MLE} = max(x_1, x_2, ..., x_n)$.

## Section 2

Bayesian Posterior Derivation for the Distribution of Serious/Fatal Injuries

Assume that $X_1, ..., X_n \overset{iid}{\sim} Beta(\alpha, \beta)$, a reasonable prior distribution would be $X_1, ..., X_n \overset{iid}{\sim} Beta(1, 1)$ which is a straight line resembling our simple linear regression for our data, for the sake of this Bayesian inference we will be using the data that includes elderly people. We will take the posterior distribution to be a binomial distribution $PosteriorBinom = p^x(1-p)^{n-x}$, using Bayes Rule we multiple it by the our prior PDF which is $p^{\alpha-1}(1-p)^{\beta-1}$ to get

$$p^x(1-p)^{n-x} \cdot p^{\alpha-1}(1-p)^{\beta-1} = p^x \cdot p^{\alpha-1} \cdot (1-p)^{n-x} \cdot (1-p)^{\beta-1} = p^{\alpha-(1+x)} \cdot (1-p)^{\beta-(1+n-x)}$$

Which gives us the posterior beta distribution of $X \sim Beta(1 + x, 1 + n - x)$